



PATENT ABSTRACTS OF JAPAN

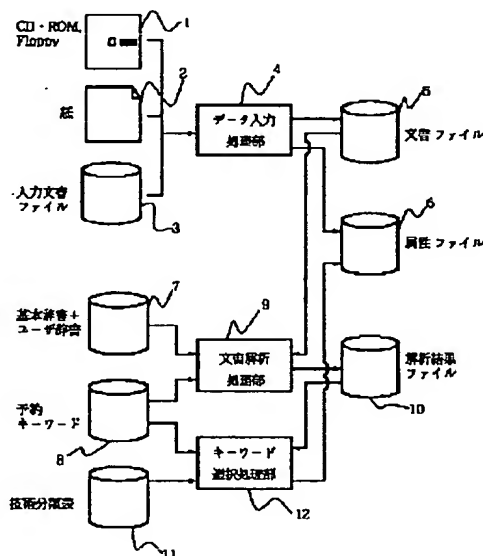
(11) Publication number: **08050593 A**(43) Date of publication of application: **20.02.96**(51) Int. Cl. **G06F 17/30**(21) Application number: **06202832**(71) Applicant: **FUJI XEROX CO LTD**(22) Date of filing: **04.08.94**(72) Inventor: **YAMAGUCHI YOSHIAKI**(54) **KEYWORD IMPARTING DEVICE**

COPYRIGHT: (C)1996,JPO

(57) Abstract:

PURPOSE: To impart keywords to a document corresponding to the various states of using keywords and document databases based on the sense of a user.

CONSTITUTION: This device is provided with a means 9 for scanning the objective document and extracting the words, and a means 9 for counting the weight based on the prescribed reference of the extracted word, a means 11 for holding the reference data with keyword including the arrangement according to the prescribed reference of two or more related words corresponding to the candidate of the keyword to be imparted and the keyword candidate, a means 12 for collating the arrangement according to the prescribed reference of the extracted word with the arrangement of the related word of the keyword imparting reference data, and a means 12 for imparting the keyword candidate corresponding to the related word to which arrangements are matched with each other as the keyword of the objective document. The candidate of the set keyword is imparted as the document keyword based on the word extracted from the document.



(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平8-50593

(43)公開日 平成8年(1996)2月20日

(51)Int.Cl.⁶

G 0 6 F 17/30

識別記号

庁内整理番号

F I

技術表示箇所

9194-5L

G 0 6 F 15/ 401

3 1 0 C

審査請求 未請求 請求項の数1 F D (全 14 頁)

(21)出願番号

特願平6-202832

(22)出願日

平成6年(1994)8月4日

(71)出願人 000005496

富士ゼロックス株式会社

東京都港区赤坂三丁目3番5号

(72)発明者 山口 義昭

神奈川県川崎市高津区坂戸3丁目2番1号

K S P R & D ビジネスパークビル

富士ゼロックス株式会社内

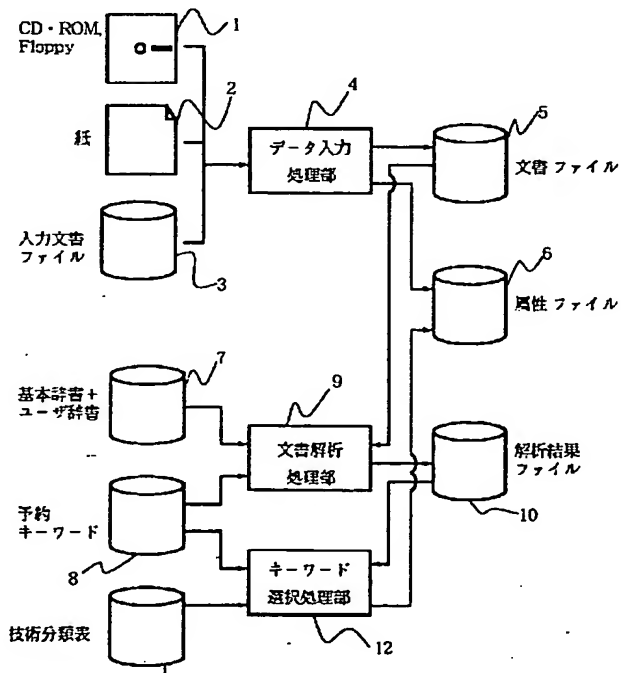
(74)代理人 弁理士 守山 辰雄

(54)【発明の名称】 キーワード付与装置

(57)【要約】

【目的】 ユーザの感覚に基づいたキーワードや文書データベースの種々な使用状況に対応したキーワードを文書に付与する。

【構成】 対象文書を走査して語を抽出する手段9と、抽出された語の所定の基準に従った重みを計数する手段9と、付与すべきキーワードの候補と該キーワード候補に対応する2以上の関連語の前記所定の基準に従った並びを含むキーワード付与基準データを保持する手段11と、抽出された語の前記所定の基準に従った並びとキーワード付与基準データの関連語の並びが一致するかを照合する手段12と、並びが一致すると判断された関連語に対応するキーワード候補を対象文書のキーワードとして付与する手段12と、を備え、任意に設定したキーワードの候補を文書から抽出した語に基づいて当該文書のキーワードとして付与する。



【特許請求の範囲】

【請求項 1】 任意に設定したキーワードの候補を文書から抽出した語に基づいて当該文書のキーワードとして付与するキーワード付与装置であって、対象文書を走査して語を抽出する手段と、前記抽出された語の所定の基準に従った重みを計数する手段と、付与すべきキーワードの候補と該キーワード候補に対応する 2 以上の関連語の前記所定の基準に従った並びを含むキーワード付与基準データを保持する手段と、前記抽出された語の前記所定の基準に従った並びと前記キーワード付与基準データの関連語の並びが一致するかを照合する手段と、前記照合する手段によって並びが一致すると判断された関連語に対応するキーワード候補を前記対象文書のキーワードとして付与する手段と、を備えたことを特徴とするキーワード付与装置。

【発明の詳細な説明】**【0001】**

【産業上の利用分野】 本発明は、任意に設定したキーワードを文書から抽出した語に基づいて当該文書のキーワードとして付与するキーワード付与装置に関する。

【0002】

【従来の技術】 文書データベース等においては、蓄積された文書データの検索に利用するために、各文書にキーワードを付与することが行われている。キーワードの付与にはキーワードの的確性、検索に際しての利便性等が要求され、これらの要求を考慮して従来より種々の方式が提案されている。

【0003】 例えば、特開平 4-243477 号公報には、キーワードを予めユーザ辞書に登録しておき、このユーザ辞書を参照して文書に自然言語処理を行ってキーワードを抽出する方式が提案されている。また、特開平 3-135669 号公報には、検索の利便性を向上させるために、文書中から抽出した語に出現回数及び出現位置の重み付けをし、重み付けの度合いの高い語を重要キーワードとして抽出する方式が提案されている。

【0004】 また、特開平 5-257979 号公報には、キーワードの的確性を向上させるために、複数のキーワード候補をグループ化して予め登録しておき、文書中からキーワード候補を抽出した際に、このグループに属するキーワード候補も自動的に抽出する方式が提案されている。また、特開平 2-28769 号公報には、キーワードの的確性を向上させるために、文書中に現れるキーワードの同義語や関連語からキーワードが表すキーワード概念を抽出し、キーワード概念の組合せを調べることにより文書全体の主題を表すキーワードを生成する方式が提案されている。

一の実際の使用状況を考察すると、機械的には生成し得ないユーザの感覚に基づいたキーワードや文書データベースの種々な使用状況に対応したキーワードを用いる方が便利の場合が多くある。例えば、技術分野の名称のように文書の内容全体を一言で表現したキーワード、担当者の名前のように或る範囲の文書を担当者毎に区別するキーワード、プロジェクトの名称のように或る範囲の文書を用途毎に区別するキーワード等のようにユーザが任意に設定したキーワードをデータベースに蓄積されている各文書に付与した方が、後のキーワード検索において便利の場合がある。また、キーワードによって分類化された各文書を更に細分類化するために、抽出されたキーワードから更に 2 次的なキーワードを抽出して各文書に付与した方が便利の場合もある。

【0006】 しかしながら、上記した従来のキーワード付与方式にあつては、文書に付与されるキーワードは、当該文書中から抽出されたキーワードから機械的に何らかの関連性があると判断された或る範囲の語でしかなく、ユーザが任意に設定した語にはなり得ないものであり、また、細分類化等のために抽出したキーワードから更に 2 次的なキーワードを抽出することもできなかった。

【0007】 本発明は、上記従来の事情に鑑みなされたもので、ユーザの感覚に基づいたキーワードや文書データベースの種々な使用状況に対応したキーワードを文書に付与することができるキーワード付与装置を提供することを目的とする。

【0008】

【課題を解決するための手段】 上記目的を達成するため、本発明のキーワード付与装置は、任意に設定したキーワード候補を文書から抽出した語に基づいて当該文書のキーワードとして付与するキーワード付与装置であつて、対象文書を走査して語を抽出する手段と、前記抽出された語の所定の基準に従った重みを計数する手段と、付与すべきキーワードの候補と該キーワード候補に対応する 2 以上の関連語の前記所定の基準に従った並びを含むキーワード付与基準データを保持する手段と、前記抽出された語の前記所定の基準に従った並びと前記キーワード付与基準データの関連語の並びが一致するかを照合する手段と、前記照合する手段によって並びが一致すると判断された関連語に対応するキーワード候補を前記対象文書のキーワードとして付与する手段とを備えたことを特徴とする。

【0009】 ここに、上記の所定の基準とは、その語の出現頻度の他、その語が出現する文書中の位置で重み付けした出現頻度等であり、要は、その語の重要度を表す基準で、データベースの使用状況等に応じて任意に設定されるものである。

書から抽出された語の所定の基準に従った並びに基づいて、当該対象文書に付与するキーワードを決定するようにしている。すなわち、ユーザが設定した任意のキーワード候補に対応させて2以上の関連語を前記と同じ基準に従って並べておき、この関連語の並びと前記文書から抽出された語の並びが一致するキーワード候補を対象文書のキーワードとして採用する。

【0011】したがって、文書から抽出した語から機械的には導き出せないユーザの感覚に基づいたキーワードや文書データベースの種々な使用状況に対応したキーワードをキーワード候補として設定し、これらキーワード候補の中から対象文書にキーワードを付与することができる。すなわち、技術分野の名称、担当者の名前、プロジェクトの名称等のように対象文書から機械的に導き出せない語もキーワードとして対象文書に付与ことができ、また、文書を更に細分類化するための2次的なキーワードを対象文書に付与することができる。

【0012】

【実施例】本発明の一実施例に係るキーワード付与装置を図面を参照して説明する。図1に示すように、本実施例のキーワード付与装置は、3つの形式の入力手段1、2、3と、入力された文書データを本装置で取り扱うことができるコードデータに変換するデータ入力処理部4と、コード変換された文書データを格納する文書ファイル手段5と、文書ファイル手段5に格納された文書の検索や分類に用いる情報を格納する属性ファイル手段6と、文書データを形態素解析するために用いる基本辞書及びユーザ辞書手段7と、形態素解析及びキーワード選択に用いる語を予約キーワードとして格納する予約キーワード格納手段8と、文書データを形態素解析する文書解析処理部9と、形態素解析されて構成語に分解された文書データを格納する解析結果ファイル手段10と、文書に付与すべきキーワードの候補と該キーワード候補に対応する2以上の関連語の所定の基準に従った並びを記述した技術分類表（キーワード付与基準データ）を格納する技術分類表格納手段11と、技術分類表に記述されたキーワード候補の中から所定の基準に該当する語をキーワードとして選択するキーワード選択処理部12とを備えている。

【0013】入力手段1からはフロッピーディスクやCD-ROM等からの文書のコード情報やイメージ情報が入力され、入力手段2からは印刷物等から読み取った文書のイメージ情報が入力され、入力手段3からは磁気ディスクに格納された文書データやネットワークで接続された他のシステムからの文書データが入力される。

【0014】データ入力処理部4は、入力手段1、2、3から入力された文書データを、イメージデータはコードデータに変換し、更にコードデータも本装置で取り扱

た、データ入力処理部4は、この処理に際して文書名、文書の作成年月日、文書の作成者等といった格納された文書の検索や分類に便利な情報が得られる場合には、これら情報を属性ファイル手段6に格納する。属性ファイル手段6にはこのような情報の他、キーワード選択処理部12で選択されたキーワードも入力され、属性ファイル手段6はこれら情報及びキーワードを文書毎に対応付けて格納する。

【0015】基本辞書及びユーザ辞書手段7並びに予約キーワード格納手段8の内容は文書解析処理部9による文書の形態素解析に用いられ、特に、予約キーワード格納手段8には技術分類表格納手段11の関連語と一致することを意図した多数の語が格納されている。これら辞書手段7並びに予約キーワード格納手段8の内容は最適な形態素解析と語の抽出が行われるように、ユーザーによって予め作成及び編集されている。文書解析処理部9は、辞書手段7並びに予約キーワード格納手段8を参照して、文書ファイル手段5に格納された文書データを文書毎に形態素解析し、この解析結果を文書毎に解析結果ファイル手段10に格納する。

【0016】技術分類表格納手段11はキーワード付与基準データとしての技術分類表を格納しており、この技術分類表には文書に付与すべき多数のキーワード、および、これらキーワードにそれぞれ対応した複数の関連語が記述されている。すなわち、技術分類表は図2に示すような構成となっており、技術分類の欄には文書に付与すべきキーワードとして「ファイル検索」、「ファイル管理」、「ディスク接続」等といった語が記述され、重要度順位の欄には各語に対応させて「ファイル、検索、・・・」、「ファイル、管理、・・・」、「Disk、制御、・・・」等といった複数の関連語が記述されている。

【0017】本実施例では所定の基準として語の文書中における出現頻度を採用しており、各重要度順位の欄における関連語はこの出現頻度の高い順に左から配列されている。例えば、付与すべきキーワード「ファイル検索」に対応して関連語「ファイル、検索、・・・」が出現頻度の高い順に並べて記述されている。この技術分類表の技術分類の欄及び重要度順位の欄に記述される語はユーザーによって任意に設定されるのもであり、本実施例では、文書を技術分野毎に分類することを意図しているため、文書に付与すべきキーワードを記述分野を示す語として技術分類の欄に記述してある。

【0018】キーワード選択処理部12は解析結果ファイル手段10に格納されている構成語に分解された文書データを文書毎に処理するものであり、この文書データの中から予約キーワード格納手段8に登録されている語を抽出し、抽出された語の当該文書中における出現回数を

関連の語並びに照合し、抽出された語の並びが関連語の並びに一致する技術分類の欄の語を文書のキーワードとして選択して、属性ファイル手段6に当該文書のキーワードとして登録する。

【0019】上記構成のキーワード付与装置における処理を図3乃至図6に示すフローチャートを参照して説明する。まず、キーワード付与装置への文書データの入力処理は、データ入力処理部4によって図3に示す手順で行われる。すなわち、入力手段1、2、3から入力された文書データを読み取り（ステップS1）、この文書データを必要なコード変換を行って文書ファイル手段5に格納するとともに（ステップS2）、文書データから抽出した文書名等の属性データを属性ファイル手段6に格納する（ステップS3）。この属性ファイル手段6には文書毎に属性ファイルが形成され、各文書に対応して属性データが格納される。

【0020】そして、入力手段1、2、3からは各手段に入力された文書数を示す信号や入力された全ての文書の出力終了を示す信号がデータ入力処理部4に入力されており、データ入力処理部4はこの信号に基づいて全ての文書データの入力処理が終了したかを判断し（ステップS4）、全ての文書データの入力処理が終了するまで上記ステップS1乃至S3の処理を繰り返し行う。

【0021】次いで、上記のようにして入力された文書データの解析処理は、文書解析処理部9によって図4に示す手順で行われる。すなわち、文書ファイル手段5から1つの文書のデータを読み取り（ステップS11）、基本辞書及びユーザ辞書手段7並びに予約キーワード格納手段8の内容を参照して、この文書データを走査しつつ形態素解析して構成語に分解し（ステップS12）、構成語に分解された文書データを解析結果ファイル10に格納する（ステップS13）。この分解された文書データは構成語を区切コードによって区切った構造であり、各文書毎に対応付けたファイルとして解析結果ファイル10に格納される。

【0022】そして、上記データ入力処理部4で入力した文書数と実際に解析処理が終了した文書数とから、文書ファイル手段5に格納された全ての文書データについての解析処理が終了したかを判断し（ステップS14）、全ての文書データの解析処理が終了するまで上記ステップS11乃至S13の処理を繰り返し行う。

【0023】次いで、上記のようにして解析された文書データに基づいたキーワードの選択処理は、キーワード選択処理部12によって図5及び図6に示す手順で行われる。すなわち、解析結果ファイル手段10から1つの文書のデータを読み取り（ステップS21）、予約キーワード格納手段8に格納されている語に一致する語をこの文書データから抽出して、抽出された語の当該文書デ

とを確認した語（ステップS23）、当該抽出された語をキーワードとして属性ファイル手段6の対応する属性ファイルに格納する（ステップS24）。なお、この抽出された語によるキーワードは従来より一般的なキーワードであり、本発明に特有なキーワードは後述するステップS34で登録されるキーワードである。

【0024】そして、上記のような抽出語のキーワード登録が終了すると、予約キーワード格納手段8に格納されている語の数に基づいて、予約キーワード格納手段8に格納されている全ての語について文書データからの抽出処理が終了したかを判断し（ステップS25）、終了していない場合には上記ステップS22乃至S24の処理を繰り返し行う。一方、全ての予約キーワード語について文書データからの抽出処理が終了した場合には、抽出した各語の出現頻度順の並びを技術分類表格納手段11の格納された技術分類表の内容に参照する処理を行う。

【0025】まず、テーブル（1）に抽出した各語を出現頻度順に並べ（ステップS26）、図2に示した技術分類表の技術分類の欄を順次示すポインタ（1）を最上段の欄（「ファイル検索」）にセットする（ステップS27）。次いで、テーブル（1）のポインタ（2）を最上段の抽出語（すなわち、出現頻度の最も高い語）にセットし（ステップS28）、技術分類表のポインタ

（1）が示す技術分類の欄に対応する重要度順位の欄の関連語を順次示すポインタ（3）を先頭の（すなわち、出現頻度が大きいとした）関連語（「ファイル」）にセットする（ステップS29）。次いで、ポインタ（2）の示す抽出語とポインタ（3）が示す関連語とを照合し（ステップS30）、両者が一致するかを判断する（ステップS31）。すなわち、この場合には、対象の文書データから最も出現頻度が高いとして抽出された語が、図2の技術分類表の最上段の欄の「ファイル」という関連語に一致するかを判断する。

【0026】この判断の結果、一致する場合には、ポインタ（2）を1つ増加させてテーブル（1）の次の抽出語を示させるとともに、ポインタ（3）も1つ増加させて技術分類表のポインタ（1）が示す技術分類の欄に対応する重要度順位の欄の次の関連語（「検索」）を示させる（ステップS32）。この後、ポインタ（2）又はポインタ（3）の先が終了か、すなわち、増加されたポインタ（2）がテーブル（1）の最後の抽出語を示している又は増加されたポインタ（3）が技術分類表のポインタ（1）が示す技術分類の欄に対応する重要度順位の欄の最後の関連語を示しているかを判断する（ステップS33）。すなわち、技術分類表のポインタ（1）が示す技術分類の欄に対応する重要度順位の欄の全ての関連語について上記の照合が終了したかを判断する。

は、上記のステップS30以降の処理を繰り返し行う一方、上記の照合による一致がとれつつ重要度順位の欄の全ての関連語について上記の照合が終了している場合には、技術分類表のポインタ(1)が示す技術分類の語を対象の文書データのキーワードとして属性ファイル6に登録する(ステップS34)。すなわち、技術分類表の重要度順位の欄に出現頻度を基準として記述された関連語の並びに、対象の文書データから抽出された語の出現頻度順の並びとが一致した場合に、この重要度順位の欄に対応する技術分類の欄の語がキーワードとして登録される。例えば、対象の文書データから「ファイル」、「検索」、・・・が頻度の高い順に抽出された場合には、対応する「ファイル検索」という技術分類を表す語が対象文書の2次的なキーワードとして登録される。

【0028】一方、上記のステップS31で一致がとれない場合、又は、上記ステップS34による登録がなされた場合には、ポインタ(1)を1つ増加させて技術分類表の次の段についての照合を行う(ステップS35)。すなわち、現在のポインタ(1)が既に技術分類表の最下段にまで達していないことを確認した後(ステップS36)、技術分類表の「ファイル管理」の欄、「ディスク管理」の欄、・・・について順次上記のステップS28乃至S35の処理を繰り返し行う。この結果、対象としている文書データについて、抽出語の並びと関連語の並びの一致がある場合には、技術分類表の技術分類の欄に記述してある語が2次的なキーワードとして付与される。

【0029】上記した1つの文書データについてのキーワード選択処理が終了すると、解析結果ファイル手段10に格納されている全ての文書データについて上記の処理が終了したかを判断し(ステップS37)、終了していない場合には解析結果ファイル手段10に格納されている次の文書データについて上記のステップS21乃至S36の処理を繰り返し行って、各文書データに技術分類表から2次的なキーワードを選択して付与し、属性ファイル手段6に登録する。

【0030】ここで、例えば、文書から抽出された語が、「ファイル」、「検索」、「管理」の3つであったとし、或る文書Aでは「ファイル」が5回、「検索」が3回、「管理」が1回使われており、他の文書Bでは「ファイル」が5回、「検索」が1回、「管理」が3回使われていたとすると、従来の方式でキーワードを付与すると、文書A、B共に同じキーワードが付与されることとなる。これに対し、本実施例のキーワード付与装置によると、文書Aには上記3つの1次的なキーワードの他に「ファイル検索」という2次的なキーワードが付与され、文書Bには上記3つの1次的なキーワードの他に「ファイル管理」という2次的なキーワードが付与され

によると、文書に付与される1次的キーワード(抽出語)の出現頻度によって選択される2次的なキーワード(技術分類)も当該文書に付与されることとなる。このため、1次的なキーワードから機械的には導き出せないユーザの感覚に基づいた2次的なキーワードや文書データベースの種々な使用状況に対応した2次的なキーワードを1次的キーワードの重要度を加味して文書に付与することができ、文書の細分類化や後のキーワード検索において多大な利便性を実現することができる。

【0032】次に、本発明の他の一実施例に係るキーワード付与装置を図面を参照して説明する。なお、前記の実施例と同一部分には同一符号を付して重複する説明は省略する。本実施例のキーワード付与装置は、文書に対して技術分類表から自動的に付与された2次的なキーワード(技術分類)を利用して、当該文書から抽出された語を1次的なキーワードとして登録するかの判断を担当のユーザからの会話形式の指示で行うものである。

【0033】本実施例のキーワード付与装置は、図7に示すように、前記の実施例と異なる構成として、キーワード選択処理部12の代わりにキーワード候補選択部21を備える他、新たな構成として、抽出キーワードインデックスファイル手段22と、キーワード設定処理部23と、表示装置24を有した指示装置25とを備えている。

【0034】キーワード候補選択部21は、前記のキーワード選択処理部12と同様に、解析結果ファイル手段10に格納されている解析された文書データから予約キーワード格納手段8に格納されている語と一致する語を抽出し、その抽出語の当該文書における出現回数をカウントし、これら抽出語を出現頻度順に並べて技術分類表格納手段11に格納されている技術分類表(図2参照)に照合し、関連語の並びが一致する技術分類の欄の語を判別する。そして、このキーワード候補選択部21は、対象としている文書データの文書識別子(文書ID)及び判別した技術分類の欄の語と共に、当該文書データから抽出した語を1次的キーワードの候補として抽出キーワードインデックスファイル手段22に格納する。

【0035】抽出キーワードインデックスファイル手段22には図8に示すようなインデックスファイルが格納されており、上記の文書ID及び技術分類の欄の語に対応して、1次的キーワードの候補である抽出語が記述されている。例えば、本実施例では、文書ID「文書T」と「文書A」に対応して抽出語「ファイル」、「検索」及び2次的キーワード候補の技術分類として「ファイル検索」が記述され、また、文書ID「文書X」に対応して抽出語「ファイル」、「管理」及び2次的キーワード候補の技術分類として「ファイル管理」が記述される。

【0036】指示装置25はユーザが担当する技術分類

ワードインデックスファイル手段22のインデックスファイルを検索し、当該技術分類に対応する文書ID及び抽出語を探索する。そして、キーワード設定処理部23は、探索された文書IDに該当する文書データを解析結果ファイル手段10から読み出して文書を表示装置24に表示すると共に、表示した文書中で探索された抽出語を他の語とは色を変える、或いは、ハイライトする等して表示する。そして、指示装置25から表示された抽出語を登録する指示が入力された場合には、この抽出語を表示されている文書のキーワードとして属性ファイル手段6に登録する。なお、この抽出語のキーワード登録と共に、対応する技術分類の語も2次的キーワードとして属性ファイル手段6に登録してもよい。

【0037】上記構成のキーワード付与装置における処理を図9乃至図10に示すフローチャートを参照して説明する。なお、キーワード付与装置への文書データの入力処理と文書データの解析処理は前記の実施例と同様であるので説明を省略する。まず、文書データからキーワード候補を抽出してインデックスファイルを作成する処理は、キーワード候補選択処理部21によって図9に示す手順で行われる。すなわち、解析結果ファイル手段10から1つの文書データを読み取り（ステップS41）、予約キーワード格納手段8に格納されている語に一致する語をこの文書データから抽出して、抽出された語の当該文書データ中における出現回数をカウントする（ステップS42）。そして、予約キーワード格納手段8に格納されている語の数に基づいて、予約キーワード格納手段8に格納されている全ての語について文書データからの抽出処理が終了したかを判断し（ステップS43）、終了していない場合には上記ステップS42の処理を繰り返す。

【0038】一方、全ての予約キーワード語について文書データからの抽出処理が終了した場合には、前記の実施例と同様にして、抽出した各語の出現頻度順の並びを技術分類表格納手段11の格納された技術分類表の内容に順次参照し、両者の並びが一致する技術分類の語を検索する（ステップS44、S45、S46）。この結果、一致を見い出せた場合には、その技術分類の語を抽出キーワードインデックスファイル手段22に格納されているインデックスファイルの技術分類の欄に登録し（ステップS47）、技術分類表の全項目を検索しても一致が見い出せなかった場合には、インデックスファイルの技術分類の欄に「その他」に登録する（ステップS48）。

【0039】次いで、この技術分類の欄に対応させて、上記の抽出語を全てインデックスファイルの抽出語の欄に登録し（ステップS49）、更に、上記の文書データの文書IDをインデックスファイルの文書IDの欄に登

記の処理が終了したかを判断し（ステップS51）、終了していない場合には解析結果ファイル手段10に格納されている次の文書データについて上記のステップS41乃至S50の処理を繰り返し行って、各文書IDに対応したインデックスファイルを作成する。

【0040】次に、上記のようにして作成されたインデックスファイルに基づいたキーワードの設定処理は、キーワード設定処理部23によって図10に示す手順で行われる。すなわち、ユーザが自分の担当する技術分類の指定を指示装置25から入力すると（ステップS61）、この技術分類に該当する項目をインデックスファイルから1つ読み出し（ステップS62）、その項目に記述されている文書IDに基づいて解析結果ファイル手段10から対応する文書データを読み出す（ステップS63）。

【0041】次いで、インデックスファイルの対応する抽出語の欄から抽出語を1つ読み出して上記の文書データに照合し、文書データ中の一致する語を他の語とは色を変え、当該文書データによる文書を表示装置24に表示する（ステップS64）。そして、色を変えて表示された抽出語をキーワードとして登録する指示がユーザによって指示装置25から入力されたかを判断し（ステップS65）、この指示が入力された場合には、この抽出語を表示されている文書のキーワードとして属性ファイル手段6に登録する。

【0042】一方、登録する指示がなかった場合、或いは、上記の登録が終了した後には、インデックスファイルの対応する抽出語の欄に記述された全ての抽出語について登録するか否かの指示を受けたかを判断し（ステップS67）、抽出語の全てについての指示を受けていない場合には、対応する抽出語の欄に記述された次の抽出語について上記のステップS64乃至S66の処理を繰り返す。そして、対応する全ての抽出語について登録するか否かの指示を受けて、これら抽出語についての処理が終了した場合には、ユーザから入力された技術分類についてインデックスファイルに登録されている全ての項目の処理が終了したかを、インデックスファイルの登録件数から判断する（ステップS68）。この結果、終了していない場合には、入力された技術分類のインデックスファイル中の次の項目（すなわち、次の文書）について上記のステップS62以降の処理を繰り返す。

【0043】例えば、ユーザから「ファイル検索」という技術分類が入力されると、図8に示すインデックスファイルに基づいて、文書Tが表示装置に表示され、当該文書T中の「ファイル」という抽出語がキーワード候補として色を変えて表示される。そして、ユーザからキーワードとして登録する指示があった場合には、この抽出

同様な処理がなされ、ユーザによる指示に基づいてキーワードとして登録される。このように、インデックスファイルの文書Tの項目についての処理が終了すると、次の文書Aの項目についても上記と同等な処理がなされ、「ファイル」、「検索」という抽出語がユーザの指示に基づいてキーワードとして登録される。

【0044】すなわち、本実施例のキーワード付与装置によると、文書から抽出した語を当該文書のキーワードとして登録するか否かの判断を、技術分類表に登録してある2次的なキーワードに基づいて割り当てた専門的なユーザに担当させることができ、正確なキーワードの付与を効率よく行うことができる。

【0045】上記した各実施例ではキーワード付与基準データを技術分類表として2次的キーワードの候補を技術分野を表す語としたため、例えば、特許公報の調査のように技術分野毎に専門的な知識を必要とする作業において、最適な担当者を割り当てて、データベースを用いた効率的な作業を実現することができる。なお、本発明はこれに限らず、図11に示すように2次的キーワードの候補を文書を担当することを割り当てられた担当者名とする等、ユーザの感覚に基づいた最適な語や文書データベースの種々な使用状況に対応した語とすることができる。

【0046】また、上記した各実施例では抽出語及び関連語の並びと決める基準を文書中での出現頻度としたが、本発明では必要に応じて種々な基準を適用することができる。例えば、文書中での語の出現位置で重み付けすれば、特許公報の調査においては特許請求の範囲の語は重くその他は軽くして、より使用状況やユーザーの感覚にあったキーワードを各特許公報に付与することができる。

【0047】また、抽出語の並びと関連語の並びとを照合方法には、上記の各実施例で示したように全ての語の並びが一致する場合のみならず、本発明では一部の語の並びが一致する場合でも2次的キーワードの登録をするようにしてもよい。例えば、抽出語の並びがA、B、C、Dで、関連語の並びがA、B、Dあった場合に、A、B、Dという一部の語の並びから、これらが一致するものとして処理したり、或いは、抽出語の並びがA、B、C、Dで、関連語の並びがA、B、C、E、Dあった場合に、A、B、C、Dという一部の語の並びから、これらが一致するものとして処理するようにしてもよい。

【0048】

【発明の効果】以上詳細に説明したように、本発明のキーワード付与装置によると、任意に設定したキーワード

の候補と該キーワード候補に対応する2以上の関連語をキーワード付与基準データとして保持しておき、対象文書から抽出された語の並びと関連語の並びが一致するかを照合し、これらの並びが一致する場合に関連語に対応するキーワード候補を対象文書のキーワードとして付与するようにしたため、文書から抽出した語から機械的には導き出せないユーザの感覚に基づいた語や文書データベースの種々な使用状況に対応した語を対象文書に2次的なキーワードとして付与することができる。このため、データベースとして蓄積された文書のキーワード検索を効率的に行うことができ、また、各文書を細分類化して能率的な文書データベースを構築することができるという効果を奏する。

【図面の簡単な説明】

【図1】本発明の一実施例に係るキーワード付与装置の構成を示すブロック図である。

【図2】本発明の一実施例に係る技術分類表の構成を示す概念図である。

【図3】本発明の一実施例に係るデータ入力処理の手順を示すフローチャートである。

【図4】本発明の一実施例に係る文書解析処理の手順を示すフローチャートである。

【図5】本発明の一実施例に係るキーワード選択処理の手順を示すフローチャートである。

【図6】本発明の一実施例に係るキーワード選択処理の手順を示すフローチャートである。

【図7】本発明の他の一実施例に係るキーワード付与装置の構成を示すブロック図である。

【図8】本発明の他の一実施例に係るインデックスファイルの構成を示す概念図である。

【図9】本発明の他の一実施例に係るキーワード選択候補処理の手順を示すフローチャートである。

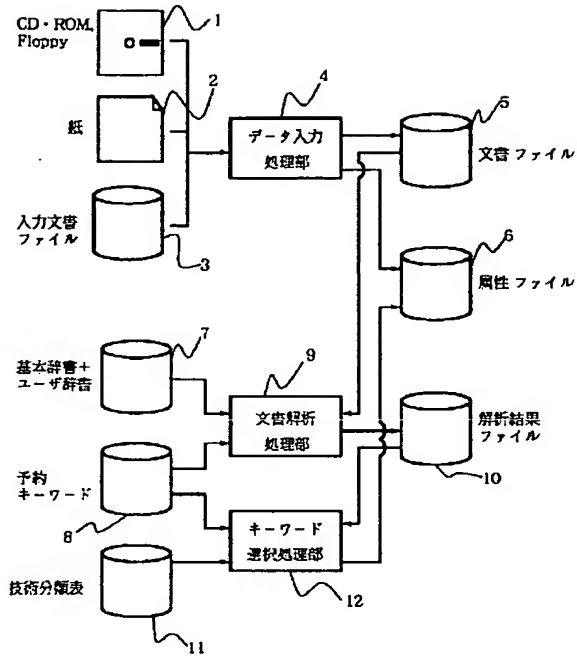
【図10】本発明の他の一実施例に係るキーワード設定処理の手順を示すフローチャートである。

【図11】技術分類表の他の一例の構成を示す概念図である。

【符号の説明】

- 4 データ入力処理部
- 5 文書ファイル手段
- 6 属性ファイル手段
- 8 予約キーワード格納手段
- 9 文書解析処理部
- 10 解析結果ファイル手段
- 11 技術分類表格納手段（キーワード付与基準データ）
- 12 キーワード選択処理部

【図 1】



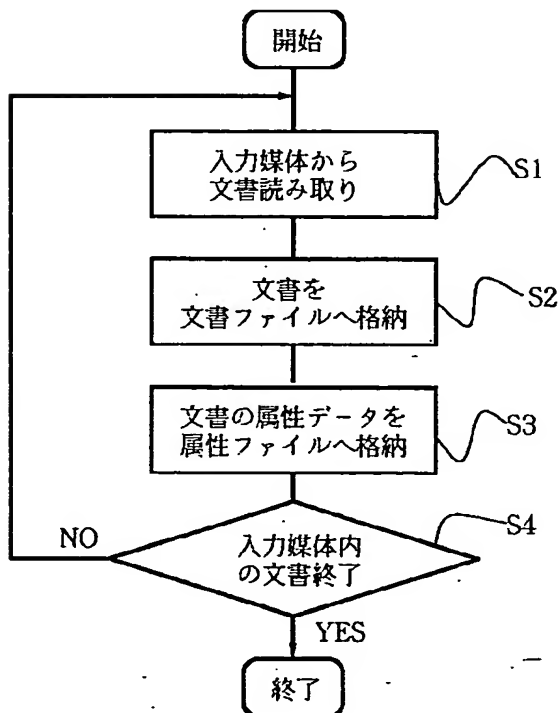
【図 2】

技術分類	重要度順位
ファイル検索	ファイル、検索、……
ファイル管理	ファイル、管理、……
ディスク検索	Disk、制御、……

【図 11】

担当者	重要度順位
山口義昭	ファイル、検索
富士太郎	ファイル、管理
赤坂次郎	Disk、制御

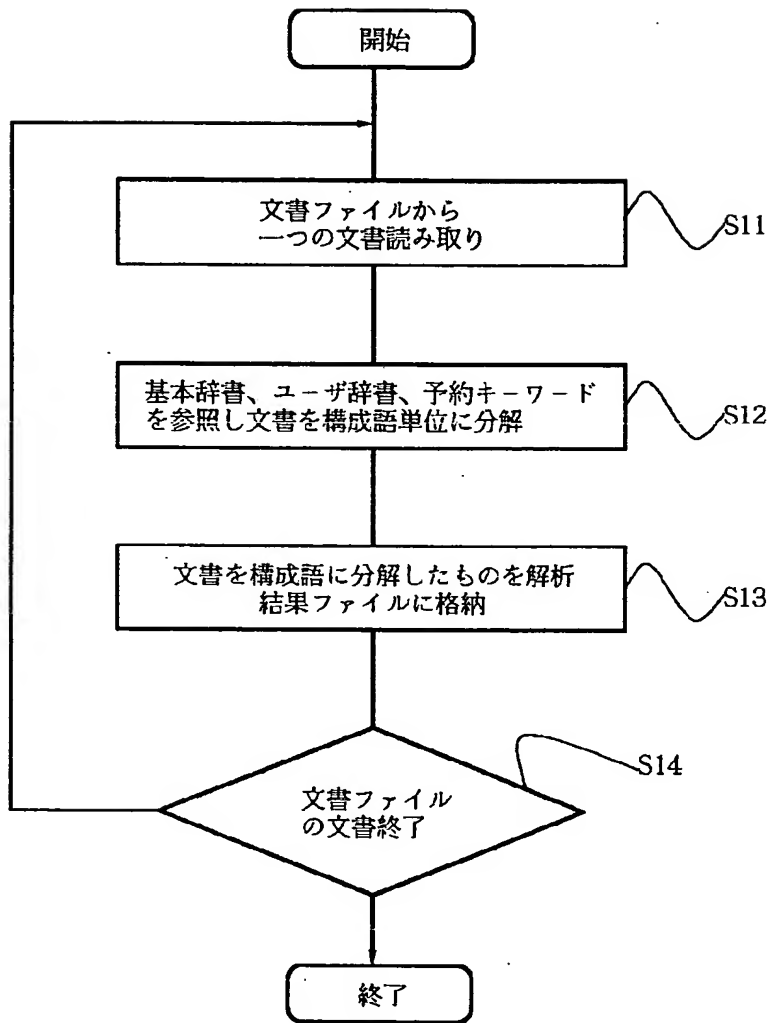
【図 3】



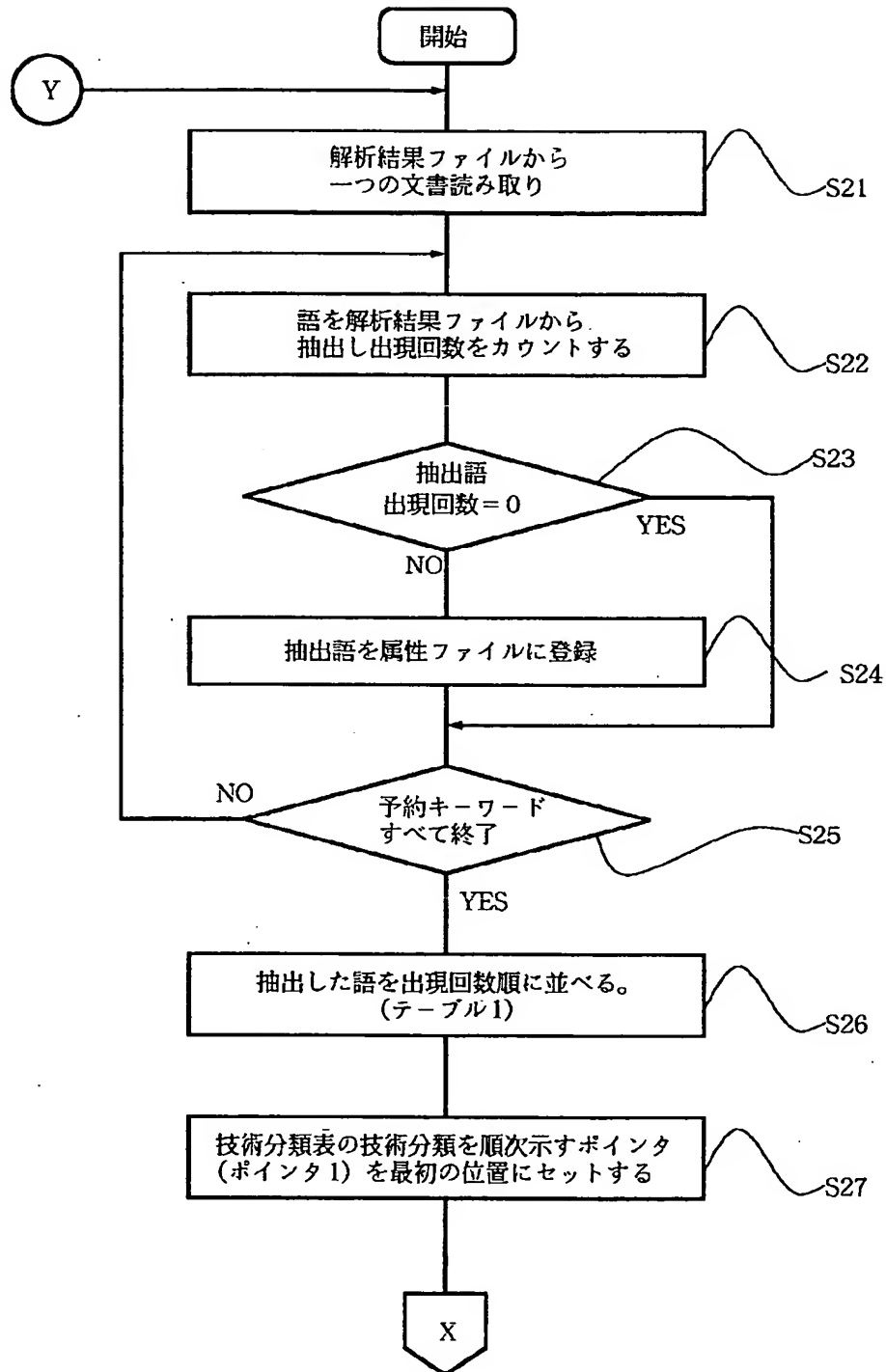
【図 8】

技術分類	抽出語	文書ID
ファイル検索	ファイル、検索	文書T
ファイル検索	ファイル、検索	文書A
⋮	⋮	⋮
ファイル管理	ファイル、管理	文書X
⋮	⋮	⋮
ディスク検索	Disk、制御	文書H

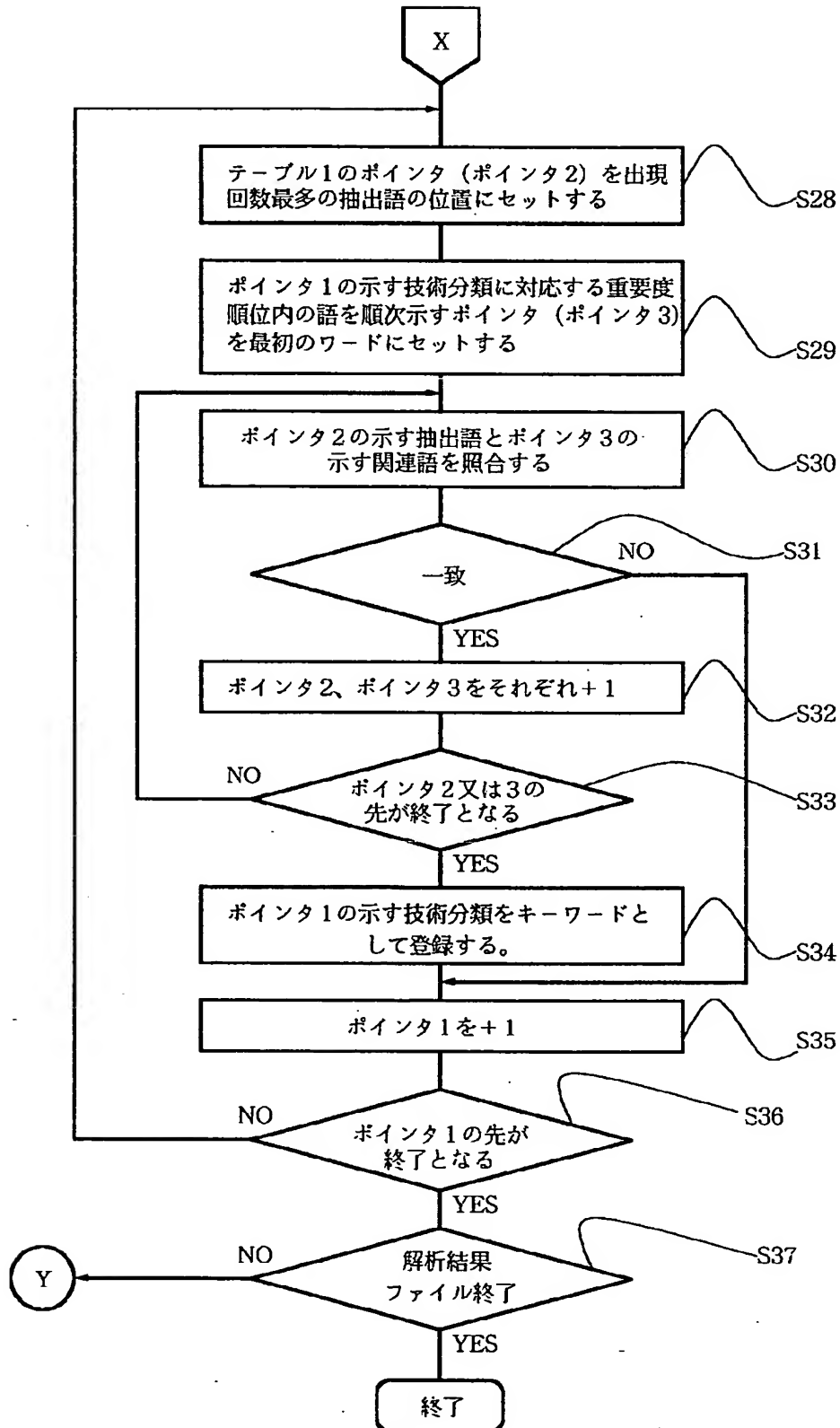
【図 4】



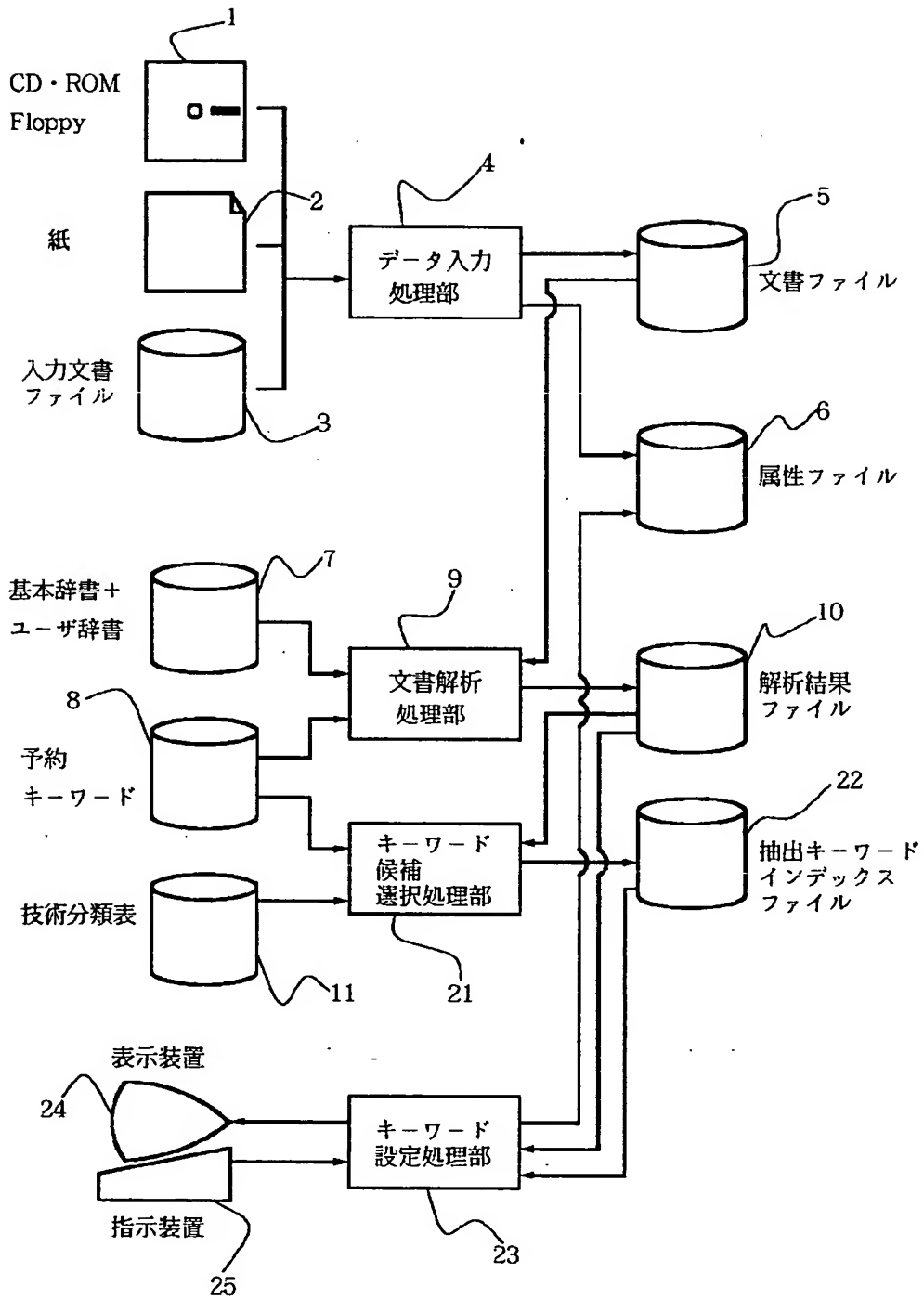
【図 5】



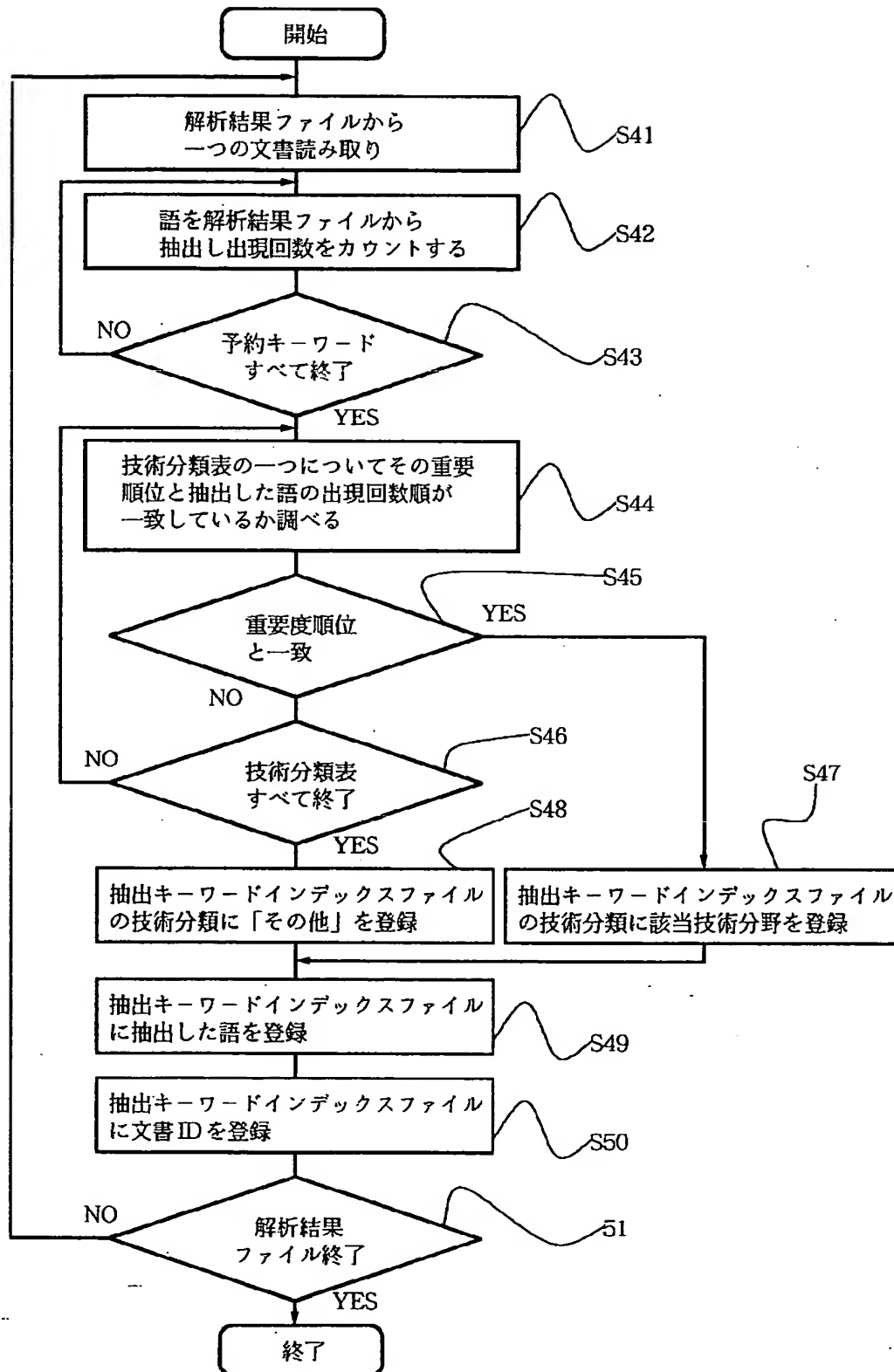
【図6】



【図 7】



【図 9】



【図 10】

